Portable Language-Independent Adaptive Translation from OCR

Quarterly R&D Status Report No. 5

Contractor:	**BBN Technologies** 10 Moulton Street, Cambridge, MA 02138
Principal Investigator:	Prem Natarajan Tel: 617-873-5472 Fax: 617-873-2473 Email: pnataraj@bbn.com
Reporting Period:	1 October 2008 – 31 December 2008

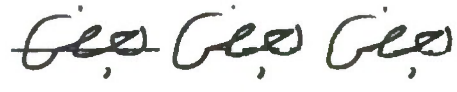
This material is based upon work supported by the Defense Advanced Research Projects Agency
DARPA/IPTO
Portable Language-Independent Adaptive Translation from OCR
MADCAT Program
ARPA Order No: X 103
Program Code: 7M30
Issues by DARPA/CMO under Contract #HR001-08-C-0004

Executive Summary

This is the fifth R&D quarterly progress report of the BBN-led team under DARPA's MADCAT program. The report is organized by technical task area.

1. Pre-Processing and Image Enhancement [BBN, Argon, Polar Rain, UMD, SUNY]

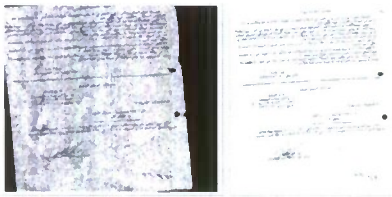
MRF-based Ruled-line Removal [BBN]: In the previous quarter, we had reported that the recognition accuracy on images with ruled-lines is significantly worse than images without ruled lines. We also investigated several approaches for ruled-line removal, but found that most approaches introduce artifacts such as breaks in glyph thereby not resulting in any significant improvement in word error rate (WER). This quarter, we developed a ruled-line removal and restoration algorithm using Markov Random Field (MRF). In our approach, a binarized image is modeled as the output of an MRF and the pixels associated with the ruled-line are restored using the belief propagation algorithm. As shown in Figure 1, our approach removes the ruled-line while still preserving the smooth edges in the handwritten glyphs. The MRF approach is also visually better than a heuristic-based approach for restoration.



(a) Input (b) Heuristic (c) MRF

Figure 1: MRF based ruled-line removal.

Morphology-based Image Enhancement [SUNY]: This quarter, we developed a noise removal and image enhancement algorithm based on a combination of morphology-based and SUNY's region growing binary image enhancement algorithm. An example is shown in Figure 2.



Degraded Image **Enhanced Image**

Figure 2: Example of Image Enhancement.

2. Page Segmentation [BBN, Argon, Lehigh, Polar Rain, UMD, SUNY]

Baseline Detection and Slant Correction [Argon]: We implemented a technique for detecting and correcting curved and slanted baselines. First, we skeletonize and filter connected components that are not associated with the baseline. Next, a sliding window encompassing at least three baseline related components is used to get sufficient context for each sub-word. Finally, local minima and junction points are detected, and Random Sample Consensus (RANSAC) algorithm is used to fit a baseline to the original skeleton. For slant correction, each endpoint above the baseline is traced back to the vicinity of the baseline, and RANSAC is again used to fit a straight line to the path. If more than 50% of the points on this line are returned as inliers by RANSAC, then it is considered sufficiently straight to use for slant correction.

Text Verification [Polar Rain]: We developed a fast and robust text verification algorithm based on Shape-DNA models. This algorithm can either be used as a tool for text segmentation or it can be used for identifying homogeneous image regions and for detecting if homogeneous regions of interest are non-text or text regions. The algorithm can also be used to identify the text type, i.e., printed or handwritten. In our approach, we first project input image segment onto a database of Shape-DNA patterns. Next, we compute histogram statistics of projection distances. Histogram of projection distances for text and non-text images as well as printed and handwritten text have different characteristics, and this feature is exploited in classifying the text images. Our approach takes only 0.5 sec for processing a 1200x800 pixel image.

3. Text Recognition [BBN, Argon, Columbia, SUNY]

Improvements in HMM based Handwriting Recognition [BBN]: In this quarter, we continued to explore techniques to reduce the WER, and also worked with new MADCAT data released by LDC.

Training Glyph Models with Additional Data: In this quarter, we updated the glyph models with 1488 images by 22 different authors released by LDC. With the addition of this set, the total amount of available training data for the text recognition system is 9741 images by 71 unique authors. Figure 3 shows the comparison of %WER separately on authors in training and authors not seen in training with different amounts of training data. The results in Figure 3 are with un-adapted decoding using PACE (percentile, angle, correlation, energy) features. Note that increasing the amount of data improves performance on the pages written by authors who are not seen in training. However, this is not the case for authors represented well in training.

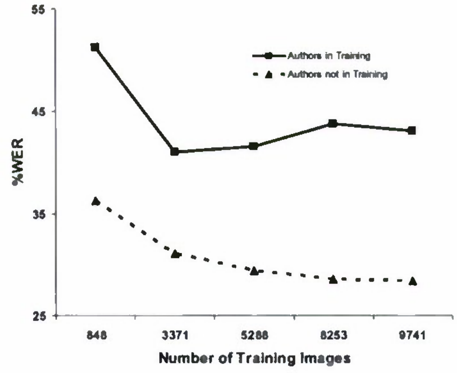


Figure 3: Comparison of recognition performance with amounts of training data.

As in the last reporting period, we trained two different OCR systems with all available training data: one with PACE features and the second with GC+PACE (gradient and concavity features in addition to PACE) features. The n-best list from decoding was re-ranked using a combination of the acoustic scores, and a language model score which does not model the "white space" token. The top best hypothesis from the re-ranked n-best list is then used to adapt the means and variances of the HMM model via MLLR estimation. A trigram language model trained on 90 million words of the GALE corpus was used for recognition. The decoding lexicon consisted on 92,000 of the most frequent words in the GALE corpus. Table 1 shows the results on the internal validation set described in the last quarter. The WER was measured by detaching punctuations and sequences of digits from other words to which they may be attached. As shown in Table 1, the GC+PACE features result in a significant improvement in WER over the standard PACE features.

System	%WER
PACE Features	40.1
GC+PACE Features	36.8

Table 1: Summary of text recognition improvements on test set.

MRF Rule-line Removal for Text Recognition: This quarter, we trained our HMM based text recognition system with MRF ruled-line removal and restoration algorithm applied to all images. Next, we decoded the test set with the models trained using GC+PACE features. Using the MRF ruled-line removal resulted in a modest improvement of 0.6% in the WER.

Stochastic Segment Modeling [BBN]: Stochastic segment modeling involves a novel combination of HMMs and 2-D matching approaches such as the bipartite graph matching (BGM). It aims to improve the HMM based handwritten Arabic text recognition by integrating long-span segment level information with the shorter-span frame based information from the HMM. In our current approach, character HMMs that use PACE features are used to force-align training transcriptions

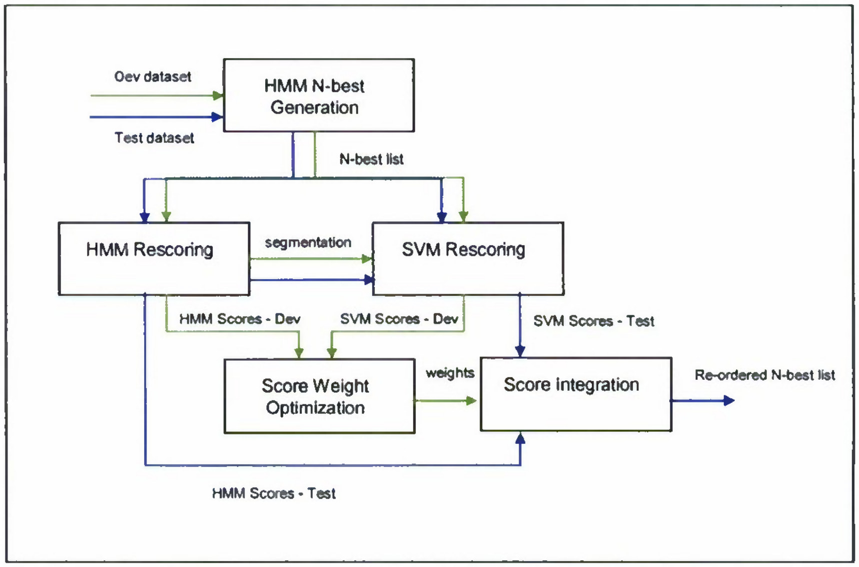


Figure 4: N-best rescoring procedure for using stochastic segment models.

to word or line images to automatically generate character boundaries. Next, 2-D images (the stochastic segments) are extracted for each character using these approximate boundaries. Features computed on these 2-D "whole character" images are used to train "segment models". In our current approach, we use support vector machines trained with GC features for modeling of stochastic character images. During recognition, the HMM character models are similarly used to generate segment boundaries for each hypothesis in the n-best list. Each segment is evaluated against the segment models which assigns a probabilistic score. Finally, the complete n-best list is rescored by combining the segment model scores with the existing HMM scores, as well as language model scores using weights that are optimized to minimize overall error rate on a development set. Figure 4 is a schematic representation of our current stochastic segment modeling approach. In Table 2, we report on improvements in WER for rescoring n-best lists on the AMA test with the above approach. As shown, using the HMM and the SVM segment scores result in a 2.3% absolute reduction in WER over using only the HMM for rescoring the n-best.

Rescoring Procedure	%WER
HMM only	55.1
HMM + SVM	52.8

Table 2: Stochastic segment based rescoring on AMA test set.

4. Integration with GALE Machine Translation [BBN]

MT on Oracle OCR Hypothesis [BBN]: Presently, we perform machine translation (MT) on the single-best OCR output. Since the 1-best OCR output has a high error rate and a lattice or n-best is likely to contain the correct answer, we performed an experiment to establish the lower bound for TER by using the best/oracle answer in the OCR n-best as the input to the MT system. As shown in Table 3, for the Devtest Part1a released by LDC, the improvement in translation error rate (TER) for using the oracle n-best hypothesis is modest. Since the oracle hypothesis has a relatively high error rate, we will repeat this experiment with a larger n-best list.

System	%WER	TER
Error-free text	-	56.4
1-best OCR hypothesis	31.5	65.8
Oracle OCR hypothesis	23.3	63.7

Table 3: Impact of using Oracle n-best hypotheses for translation.

5. Evaluation System [BBN]

In the previous quarterly report, we had discussed the design of our Phase I evaluation system. This period, we ran the Phase I evaluation system on the evaluation data provided by NIST. In addition to the primary evaluation task of running our system on handwritten images, we submitted the output of our MT engine on the reference Arabic transcriptions for the images in the evaluation test set. MT output on the reference transcriptions was used by NIST for the so-called "contrastive" condition which is designed to assess the degradation in MT performance that results from using OCR output instead of human transcriptions.

6. Project Meetings

BBN participated in four MADCAT Program Meetings during this quarter, including the following:

1. A 1-day Internal Program Review (IPR) with each sub-contractor presenting their work during Phase 1.
2. Site visit by the DARPA Program Manager, Dr. Joseph Olive. At the site visit we presented a review of the technical accomplishments of the BBN Team during Phase 1 of MADCAT.
3. Two data planning meetings to discuss data collection plans and requirements for Phase 2. The first meeting on 4 December 2008 was hosted at NIST's Gaithersburg facility, and the second on 16 December 2008, was hosted at BBN's Rosslyn facility.